# Прогнозування відтоку абонентів

Виконали:

Ярінко Богдан

Тернавський Олег

Вершинін Андрій

Датасет: <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>

Зміст

[Прогнозування відтоку абонентів 1](#_Toc183563783)

[Обробка даних 1](#_Toc183563784)

[Робота з датасетом 3](#_Toc183563785)

[Визначення спектру ризиків 5](#_Toc183563786)

[Побудова карти ризику 5](#_Toc183563787)

[Визначення рівнів ризику 6](#_Toc183563788)

[Визначення основних характеристик ризику 7](#_Toc183563789)

[Розробка матриці ризиків та втрат 8](#_Toc183563790)

[Обрання відповідних методів для прогнозування 10](#_Toc183563791)

[Ймовірність відтоку абонентів 11](#_Toc183563792)

[Покращуємо якість розроблених моделе 13](#_Toc183563793)

[Прогнозування часу відтоку абонентів 15](#_Toc183563794)

[Розрахунок можливих втрат внаслідок сукупної реалізації 17](#_Toc183563795)

[Висновки 19](#_Toc183563796)

[Антикризова стратегія 20](#_Toc183563797)

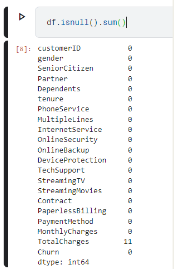
[Висновок 21](#_Toc183563798)

Сучасні компанії стикаються з викликами утримання клієнтів, що безпосередньо впливає на прибутковість. Даний проект має на меті аналіз даних про абонентів, щоб виявити закономірності, які впливають на відтік клієнтів, і розробити програми їх утримання.

Обраний набір даних, представлений IBM, містить інформацію про абонентів, які залишили компанію впродовж останнього місяця, а також характеристики їхніх рахунків, наявні послуги та демографічні дані. Аналіз цих даних дозволяє зосередитися на основних чинниках, які впливають на рішення клієнтів залишитися або піти. Проект передбачає прогнозування часу відтоку абонентів, оцінку можливих фінансових втрат та визначення ймовірності відтоку конкретних клієнтів. Це допоможе компаніям формувати ефективні стратегії для утримання своїх споживачів.

## Обробка даних

Перший крок у підготовці даних — це перевірка на наявність пропущених значень у датасеті. Деякі змінні можуть містити відсутні або некоректні значення, які потрібно відповідним чином обробити, щоб уникнути помилок у подальшому аналізі. Для цього можна скористатися наступним кодом:

****

Цей код виводить кількість пропущених значень у кожному стовпці, що дозволяє визначити, які саме змінні потребують обробки.

Наступним кроком буде обробка пропущених значень Незважаючи на те, що вони містять числові дані, вони можуть бути визначені як тип object, ймовірно через наявність пропущених або некоректних значень. Для коректної обробки необхідно привести їх до числового формату. Це можна зробити за допомогою наступного коду:



Аргумент errors='coerce' автоматично замінює некоректні значення на NaN, що полегшує їх подальшу обробку.

Пропущені дані ми заповнимо середнім значенням

Датасет містить кілька категоріальних змінних, наприклад, gender, Partner, Dependents, PhoneService, InternetService, тощо. Ці змінні необхідно перетворити в числовий формат для машинного навчання. Для цього можна використати one-hot encoding або label encoding:

Використаємо one-hot encoding для категоріальних змінних:

****

## Робота з датасетом

1. Після обробки даних ми визначили наступне:

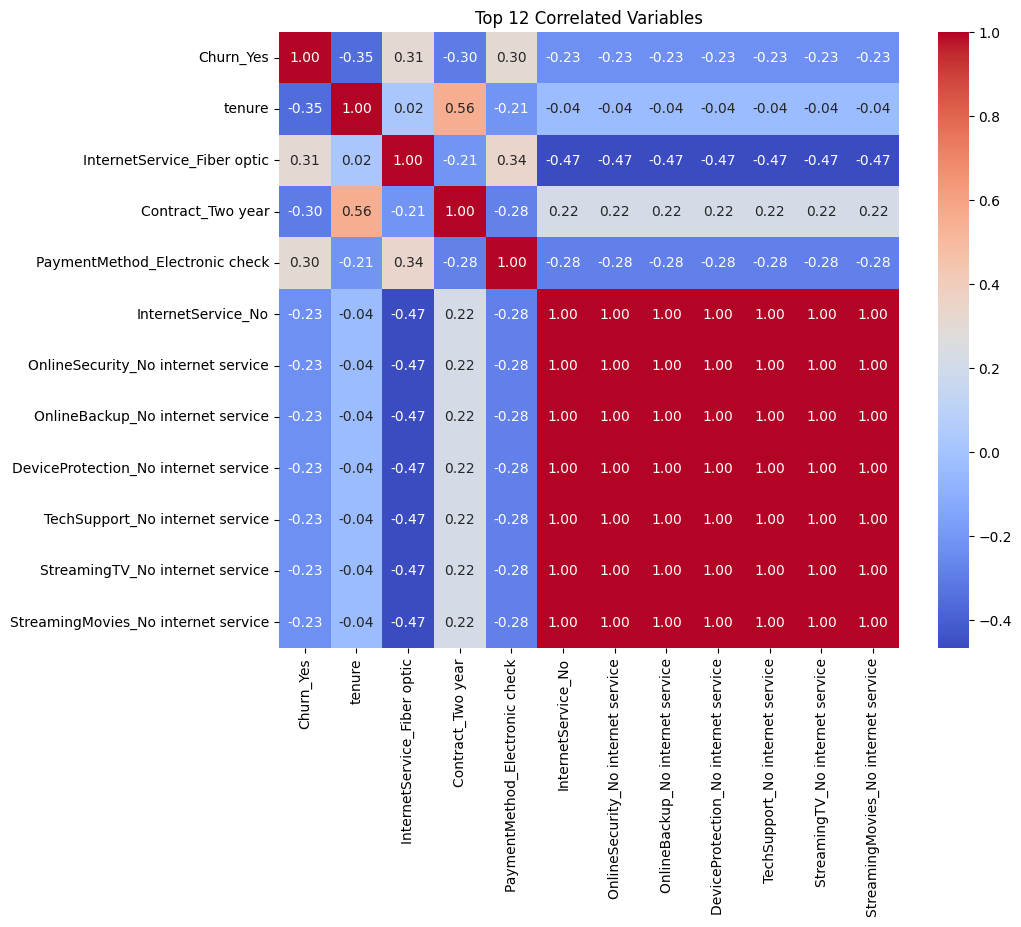
Цільова змінна в цьому датасеті — це Churn (відток абонентів), яка визначає, чи залишив клієнт компанію. Ця змінна має два значення: "Yes" або "No". У цьому випадку вона є класифікаційною змінною.

2. Причинно-наслідкові зв'язки: Деякі основні залежності:

* Churn (відтік клієнтів) як цільова змінна, яка залежить від інших факторів: тривалість користування послугами (tenure), наявність додаткових послуг, таких як OnlineSecurity, TechSupport, і загальні витрати клієнта (TotalCharges).

**Аналіз причинно-наслідкового зв’язку та визначення суттєвих змінних**

Щоб перевірити наявність зв'язку між компонентами даних, обчислимо кореляційну матрицю для кількісних змінних. Це допоможе зрозуміти, які змінні мають найбільший вплив на **Churn**:



**Основні висновки**

1. **Термін перебування (tenure)**:
   * Чим довше абонент залишається з компанією, тим менше ймовірно, що він залишить її. Кореляція -0.35.
2. **Інтернет-послуги**:
   * Абоненти, які користуються оптоволоконним Інтернетом, більш схильні до відтоку (кореляція 0.31).
   * Абоненти без Інтернет-послуг менше схильні залишати компанію (кореляція -0.23).
3. **Тип контракту**:
   * Дворічні контракти знижують ймовірність відтоку (кореляція -0.30).
4. **Спосіб оплати**:
   * Абоненти, які платять електронним чеком, мають більшу ймовірність залишити компанію (кореляція 0.30).

**Рекомендації**

* **Залучення нових клієнтів**: Розробити програми лояльності для абонентів із коротким терміном перебування.
* **Стимулювання довгострокових контрактів**: Заохочувати клієнтів підписувати дворічні контракти.
* **Аналіз способів оплати**: Розглянути зміни в обслуговуванні абонентів, які обирають електронний чек.

## Визначення спектру ризиків

 Серед фінансових ризиків можна виділити зниження доходів через високий відтік клієнтів та витрати на залучення нових абонентів. Операційні ризики пов'язані з невідповідністю послуг очікуванням клієнтів та порушеннями в обслуговуванні, що може призвести до негативних відгуків.

Стратегічні ризики включають конкуренцію з боку інших компаній та зміни в ринкових тенденціях, які можуть вплинути на попит на послуги. Репутаційні ризики виникають через негативні відгуки клієнтів та втрату довіри споживачів. Крім того, технологічні ризики пов'язані з вразливістю системи до збоїв або атак, а також невідповідністю технологій потребам абонентів.

Для управління цими ризиками важливо оцінити їхню ймовірність і вплив на бізнес, що дозволить створити карту ризиків. Зокрема, це допоможе визначити пріоритети у впровадженні заходів, спрямованих на зменшення негативних наслідків відтоку клієнтів та покращення загальної ефективності компанії.

## Побудова карти ризику

Оцінка ризиків може бути виконана за двома параметрами: ймовірність виникнення та вплив на бізнес.

* **Висока** (3)
* **Середня** (2)
* **Низька** (1)

На основі оцінок ризиків можна побудувати карту ризиків. Де у нас є такі можливі ризики



## Визначення рівнів ризику

Для управління ризиками в контексті відтоку абонентів важливо визначити три рівні ризику: допустимий, критичний і катастрофічний. Це допоможе компанії виявити, які ризики вимагають термінового реагування, а які можуть бути прийнятими в межах визначених параметрів.

**Допустимий рівень ризику**

* **Опис:** Цей рівень ризику вказує на те, що він вважається прийнятним і не потребує термінового реагування. Це може свідчити про те, що бізнес має належні стратегії для мінімізації впливу незначних ризиків.
* **Приклад:** Невеликий збій в обслуговуванні може бути звичайним явищем у бізнес-процесах, яке не викликає значних проблем. Моніторинг таких ризиків дозволяє вчасно реагувати, якщо вони почнуть наростати.
* **Необхідні дії:** Відсутність термінових дій означає, що команда може зосередитися на більш серйозних питаннях, водночас слідкуючи за розвитком ситуації.

**2. Критичний рівень ризику (А)**

* **Опис:** Цей рівень ризику вказує на те, що ситуація може загрожувати бізнесу, і вимагає контролю та стратегічного управління.
* **Приклад:** Низька лояльність нових клієнтів може бути сигналом про проблеми з обслуговуванням або пропозицією, що потребує уваги. Це ризик, який може погіршити фінансові показники компанії, якщо його не вирішити.
* **Необхідні дії:** Аналіз і розробка стратегій управління є критично важливими. Це може включати опитування клієнтів, поліпшення продуктів чи послуг, або навчання персоналу.

**3. Катастрофічний рівень ризику (Б)**

* **Опис:** Цей рівень ризику вказує на ситуації, які можуть серйозно загрожувати бізнесу і вимагають термінового реагування.
* **Приклад:** Високий відтік клієнтів може свідчити про системні проблеми в компанії, які потребують негайних змін у стратегії чи операціях.
* **Необхідні дії:** Термінові заходи, такі як розробка нових пропозицій, активне залучення клієнтів і поліпшення обслуговування, можуть допомогти зупинити негативні тенденції.

## Визначення основних характеристик ризику

Для визначення основних характеристик ризику в контексті проекту з відтоком абонентів можна сформулювати три основні задачі:

**1. Прогнозування часу (періоду) відтоку абонентів**

**Завдання:** Розробити модель для прогнозування часу, коли абоненти можуть припинити користування послугами.

**Методологія:**

* Використати методи аналізу часових рядів або машинного навчання, такі як регресійні моделі, дерева рішень або нейронні мережі.
* Визначити змінні, які можуть впливати на відтік, такі як tenure, MonthlyCharges, Contract тощо.
* Створити прогноз на основі історичних даних про відтік клієнтів.

**2. Розрахунок можливих втрат внаслідок сукупної реалізації ризику**

**Завдання:** Оцінити фінансові втрати, які можуть виникнути внаслідок відтоку абонентів.

**Методологія:**

* Визначити середній дохід від одного абонента (Average Revenue Per User, ARPU).
* Розрахувати кількість абонентів, які можуть піти за визначений період, на основі прогнозу.
* Помножити прогнозовану кількість абонентів на ARPU, щоб отримати загальні можливі втрати.

**3. Оцінювання ймовірності відтоку окремого абоненту**

**Завдання:** Визначити ймовірність того, що окремий абонент відмовиться від послуг.

**Методологія:**

* Використати статистичні моделі, такі як логістична регресія, для оцінки ймовірності відтоку на основі факторів, що впливають на рішення абонента.
* Визначити змінні, що можуть корелювати з відтоком, наприклад, Contract, PaymentMethod, OnlineSecurity, тощо.
* Провести аналіз за групами, щоб зрозуміти, які сегменти абонентів мають вищу ймовірність відтоку.

## Розробка матриці ризиків та втрат

Матриця ризиків допоможе виявити ключові ризики, які можуть вплинути на стабільність бізнесу, та визначити пріоритети в управлінні ними. Розглянемо її



У нашій матриці ризиків та втрат ми визначили ризики на основі ймовірності та впливу, пов’язаних із відтоком абонентів. Катастрофічними розглядаються ризики, такі як високий відтік клієнтів, зниження доходів через високий рівень повернень  і втрата ключових клієнтів. Ці ситуації можуть призвести до серйозних фінансових втрат і зниження репутації компанії, оскільки вони вказують на системні проблеми в обслуговуванні клієнтів та стратегічні помилки у виборі цільової аудиторії.

На рівні критичного впливу перебувають ризики, що включають негативні відгуки клієнтів, зниження лояльності нових клієнтів, низьку купівельну активність, зниження попиту на послуги. Ці ризики свідчать про незадоволеність абонентів, яка може призвести до фінансових втрат і загрожує стабільності компанії.

Допустимі ризики це невеликі збої в обслуговуванні, витрати на обслуговування та відсутність нових пропозицій. Ці фактори сигналізують про потенційні проблеми з утриманням клієнтів, які потребують уваги для запобігання подальшому погіршенню ситуації. Хоча ці ризики не є критичними, їх все ж варто враховувати для запобігання можливим ускладненням у бізнес-процесах.

## Обрання відповідних методів для прогнозування

Для прогнозування часу відтоку абонентів, можливих втрат та ймовірності відтоку важливо вибрати відповідні методи аналізу, які забезпечать точність і надійність результатів.

Прогнозування часу відтоку абонентів. Прогнозування часу відтоку абонентів може здійснюватися за допомогою моделей виживання, зокрема моделі Каплан-Мейєра. Ця модель оцінює ймовірність залишитися клієнтом на основі тривалості перебування (tenure) та бінарної змінної, що вказує на факт відтоку (1 = відтік, 0 = залишився). Візуалізація кривої виживаності дозволяє виявити патерни у даних про відтік.

Додатково, модель пропорційних ризиків Кокса (Cox Proportional Hazards Model) враховує вплив характеристик абонентів, таких як щомісячні витрати та тип контракту. Це дозволяє отримати прогноз часу до відтоку для кожного клієнта, що допомагає компаніям розробляти стратегії утримання клієнтів.

Розрахунок можливих втрат Для оцінки фінансових втрат від відтоку абонентів ми спочатку розраховуємо ймовірність відтоку для кожного абонента, що дозволяє оцінити ризик на основі їх характеристик. Потім, використовуючи цю ймовірність, обчислюємо прогнозовані втрати, враховуючи щомісячні витрати та прогнозований час до відтоку.

Метод Монте-Карло дозволяє моделювати різні сценарії, генеруючи випадкові ймовірності відтоку для кожного абонента. Після 10,000 симуляцій отримуємо розподіл можливих втрат, з якого розраховуємо середні втрати та 95% довірчий інтервал.

*Візуалізація результатів:*

Для кращого розуміння результатів побудуємо два графіки:

Гістограма розподілу симульованих втрат: Показує, як розподіляються втрати, з середніми втратами та 95% довірчим інтервалом.

Графік середніх втрат з довірчими інтервалами: Демонструє середні втрати з відповідними довірчими інтервалами.

Таким чином, поєднання моделі оцінювання ймовірності t і методу Монте-Карло забезпечує комплексний підхід до оцінки можливих фінансових втрат від відтоку абонентів.

*Оцінювання ймовірності відтоку*

Для оцінки ймовірності відтоку абонентів ми використовуємо кілька моделей машинного навчання, що забезпечує точніші прогнози. Зокрема, застосовуються:

Логістична регресія: Оцінює ймовірність відтоку на основі лінійних залежностей, проста у використанні та інтерпретації.

Випадковий ліс (Random Forest): Використовує ансамбль дерев рішень для покращення точності прогнозів, добре справляється з нелінійними залежностями.

Градієнтний бустинг (Gradient Boosting): Будує дерева послідовно, виправляючи помилки попередніх, що забезпечує високу точність.

Метод опорних векторів (Support Vector Machine): Розділяє класи у багатовимірному просторі, ефективний для складних структур даних.

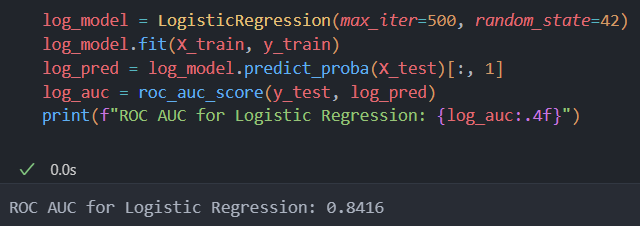
Моделі навчаються на історичних даних про абонентів, після чого використовуються для прогнозування ймовірності відтоку.

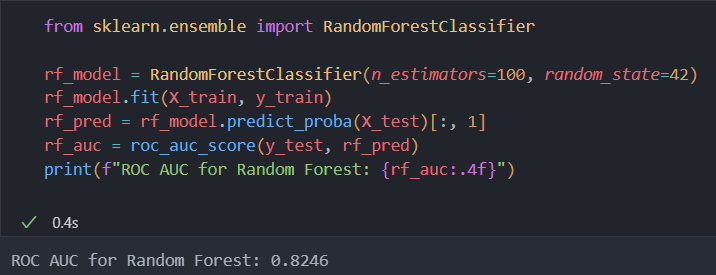
Для вибору найкращої моделі порівнюються їхні показники, такі як точність та ROC-AUC.

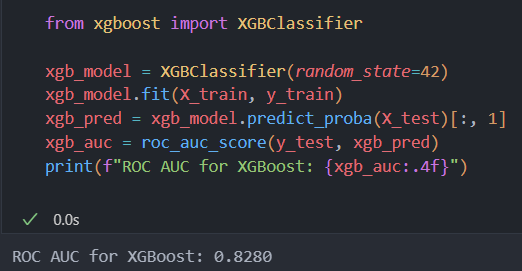
Таким чином, використання кількох моделей забезпечує комплексний підхід до оцінки ймовірності відтоку, що дозволяє отримати точні прогнози та краще розуміти ризики відтоку клієнтів.

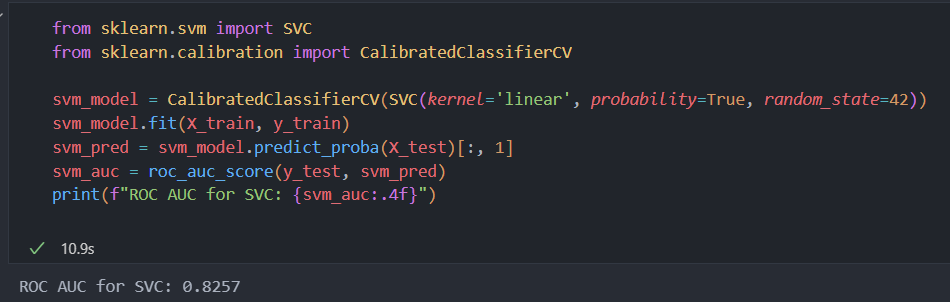
## Ймовірність відтоку абонентів

Побудуємо початкові моделі  та поглянемо на результати: Ми використали кілька популярних алгоритмів, включаючи Random Forest, XGBoost, Support Vector Machine (SVM) та логістичну регресію. Для оцінки ефективності моделей ми використовували метрики ROC AUC, точність (Accuracy), точність (Precision), повноту (Recall) та F1 Score.









Результати моделей:



* Логістична регресія показала найвищу точність (Accuracy) (0.8070) та F1 Score (0.6092), що свідчить про її загальну ефективність у класифікації.
* Random Forest має точність (Accuracy) 0.7878 та F1 Score 0.5517, що вказує на деякі проблеми з виявленням позитивних класів.
* Градієнтний бустинг показавточність  (Accuracy) 0.7899 та F1 Score 0.5795, що свідчить про схожі труднощі з виявленням позитивних випадків.
* Метод опорних векторів (SVC) досягнув  точності Accuracy 0.7949 та F1 Score 0.5693, що також вказує на потребу в покращенні.

Загалом, результати вказують на те, що логістична регресія має найкращу Accuracy та F1 Score, але всі моделі демонструють потребу в подальшому вдосконаленні для покращення виявлення позитивних класів.

## Покращуємо якість розроблених моделе

Використаємо GridSearchCV, який дозволяє знаходити найкращі гіперпараметри для покращення продуктивності класифікаторів. Ми провели оптимізацію для чотирьох моделей: Random Forest, логістичної регресії, градієнтного бустингу та Support Vector Machine (SVM).

Випадковий ліс:

Найкращий ROC AUC: 0.8439

Найкращі параметри моделі: {'max\_depth': 10, 'min\_samples\_split': 20, 'n\_estimators': 300}

Логістична регресія:

Найкращий ROC AUC: 0.8407

Найкращі параметри моделі: {'C': 10, 'solver': 'liblinear'}

Градієнтний бустинг:

Найкращий ROC AUC: 0.8465

Найкращі параметри моделі: {'learning\_rate': 0.1, 'max\_depth': 3, 'n\_estimators': 50}

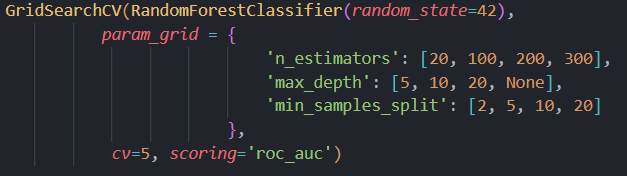
Опорні вектори:

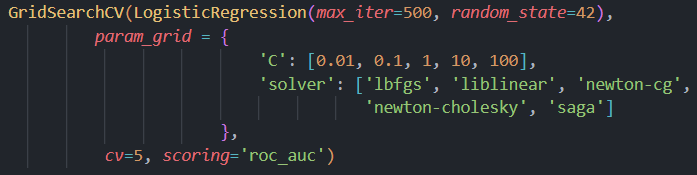
Найкращий ROC AUC: 0.8295

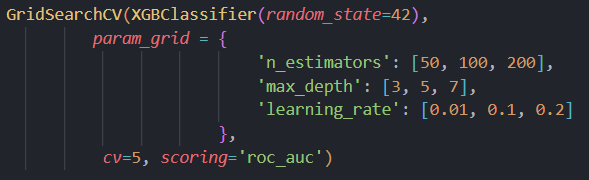
Найкращі параметри моделі: {'C': 0.01, 'gamma': 'scale'}

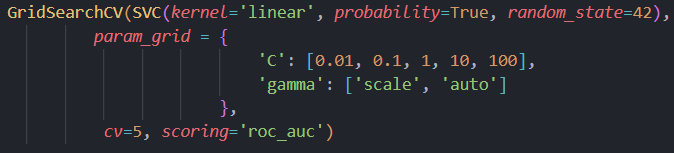
Аналіз результатів

З отриманих результатів видно, що градієнтний бустинг досяг найвищого ROC AUC (0.8465), що свідчить про його найкращу здатність розрізняти позитивні та негативні класи серед усіх моделей. випадковий ліс також показав хороші результати з ROC AUC 0.8439, а логістична регресія продемонструвала ROC AUC 0.8407.

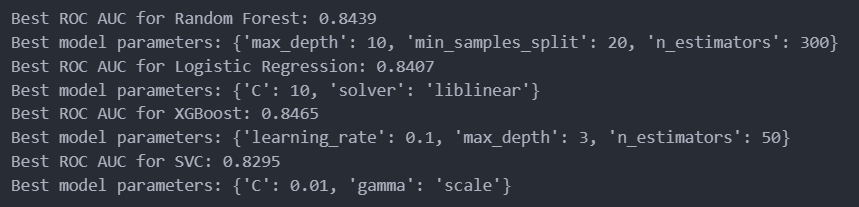




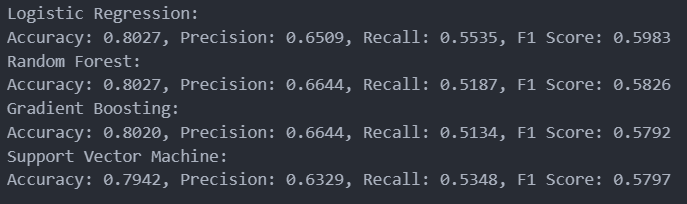




Параметри, які були оптимізовані для кожної моделі, вказують на те, що правильний вибір гіперпараметрів позитивно впливає на продуктивність



Оцінки моделей:



Загалом, логістична регресія та Random Forest показали найвищу точність, але всі моделі демонструють схожі результати, що вказує на потребу в подальшому вдосконаленні для покращення виявлення позитивних класів.

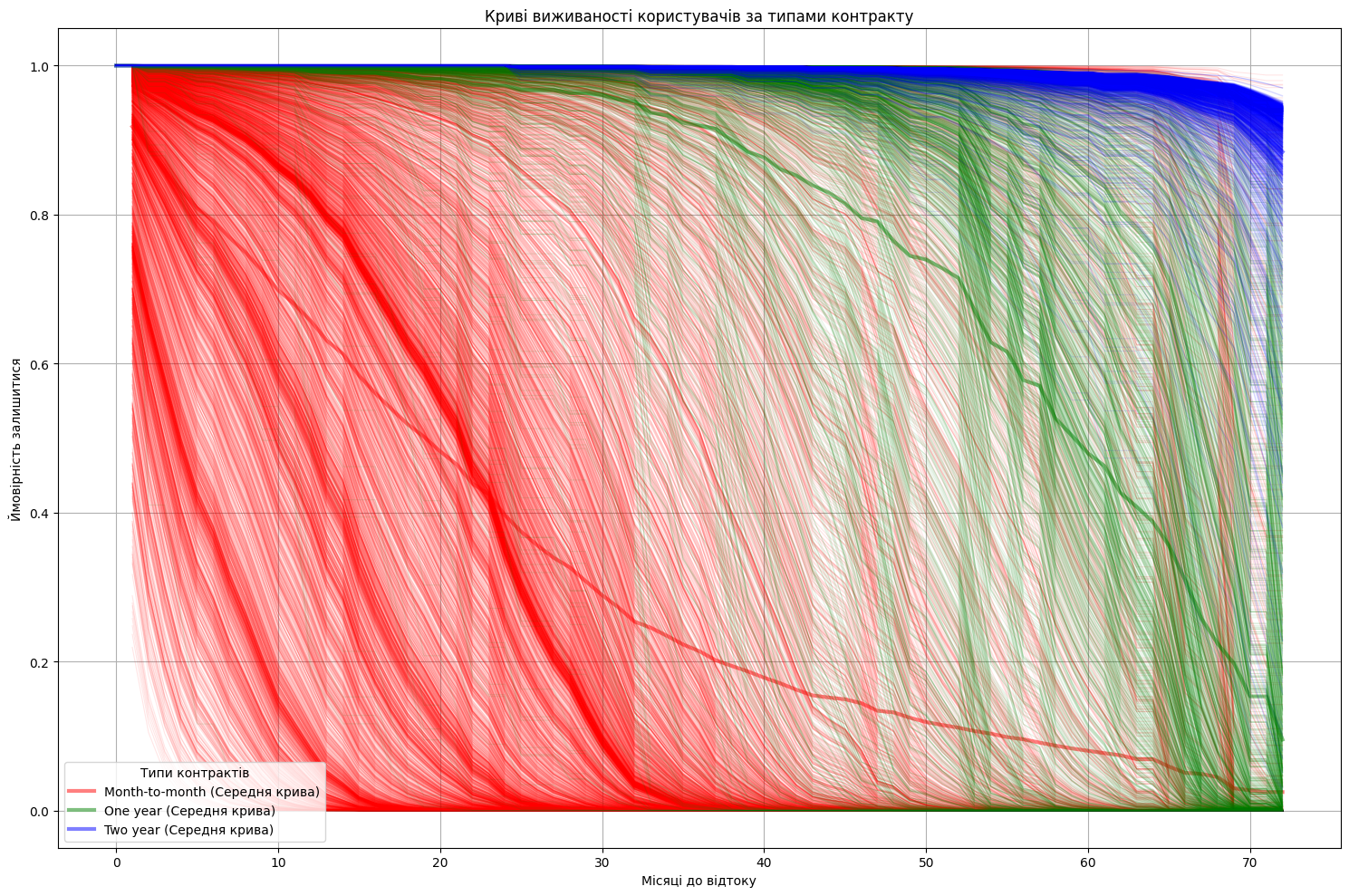
## Прогнозування часу відтоку абонентів

Для прогнозування часу до відтоку клієнтів використовується модель пропорційних ризиків Кокса. Ця модель дозволяє оцінити, як змінні, такі як щомісячні витрати, загальні витрати чи тип контракту, впливають на час до відтоку клієнта. Вона є однією з найпоширеніших методик у статистичному аналізі виживаності, оскільки враховує вплив кількох факторів одночасно.

Модель Кокса була навчена на підготовлених даних із використанням бібліотеки lifelines. Ця модель оцінює вплив змінних на час до відтоку клієнта, що дозволяє отримати такі ключові результати:

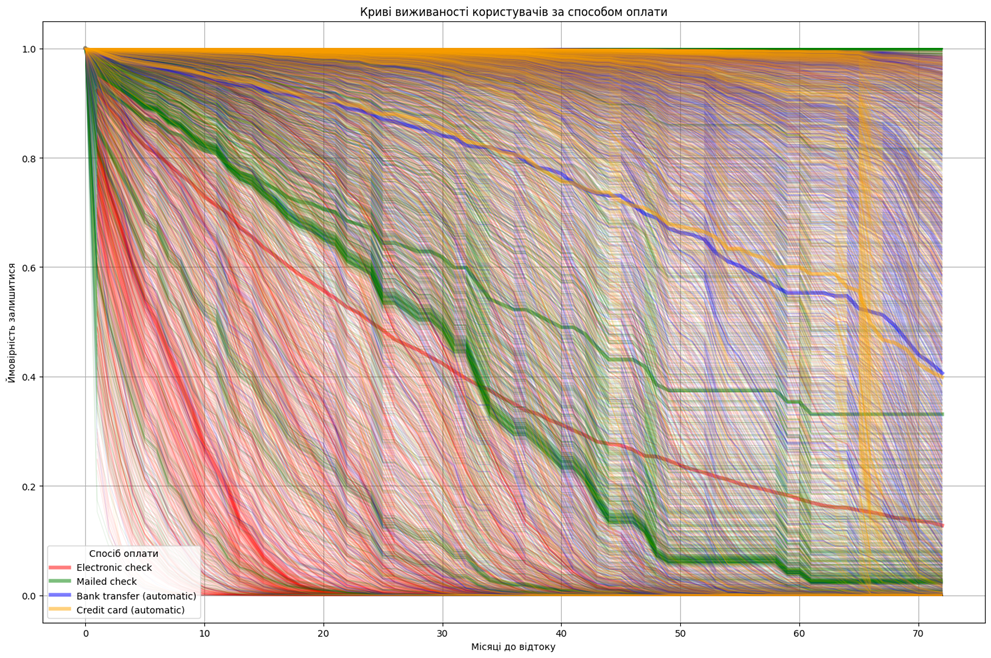
* Прогнозований час до відтоку (predicted\_time\_to\_churn).
* Ймовірність відтоку клієнта (churn\_probability).

Розглянемо тенденцію відтоку користувачів у розрізі їхньої підписки. Існує місячна, річна або дворічна підписка. побудуємо криві виживаності для кожної з груп:



з графіку видно, що користувачі, у яких місячний тип контракту, більш схильні до відтоку у перші місяці, в той час, як абоненти з дворічним контрактом найдовше затримуються в компанії

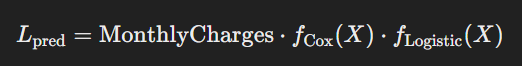
Такий же графік побудуємо за типом контракту:

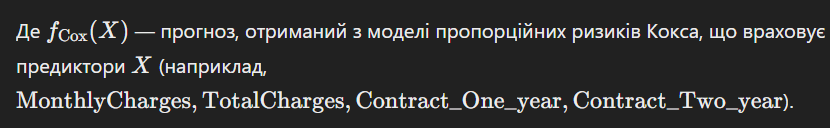


Абоненти, що оплачували електронним чеком є найбільш лояльними до компанії та залишаються найдовше з нею. А ті, хто оплачували банківським переказом, покидають її найшвидше.

Розглянемо приклад для конкретних абонентів. Побудуємо для кожного криву виживаності та отримаємо прогнозований час до відтоку за допомогою моделі Кокса. Оцінимо загальну ймовірність абонента залишити компанію Логістичною регресією з найкращими параметрами, які ми були визначили. І порахуємо втрати, які понесе компанія за умови, що клієнт її покине.

Втрати ми обраховуватимемо за наступною формулою:







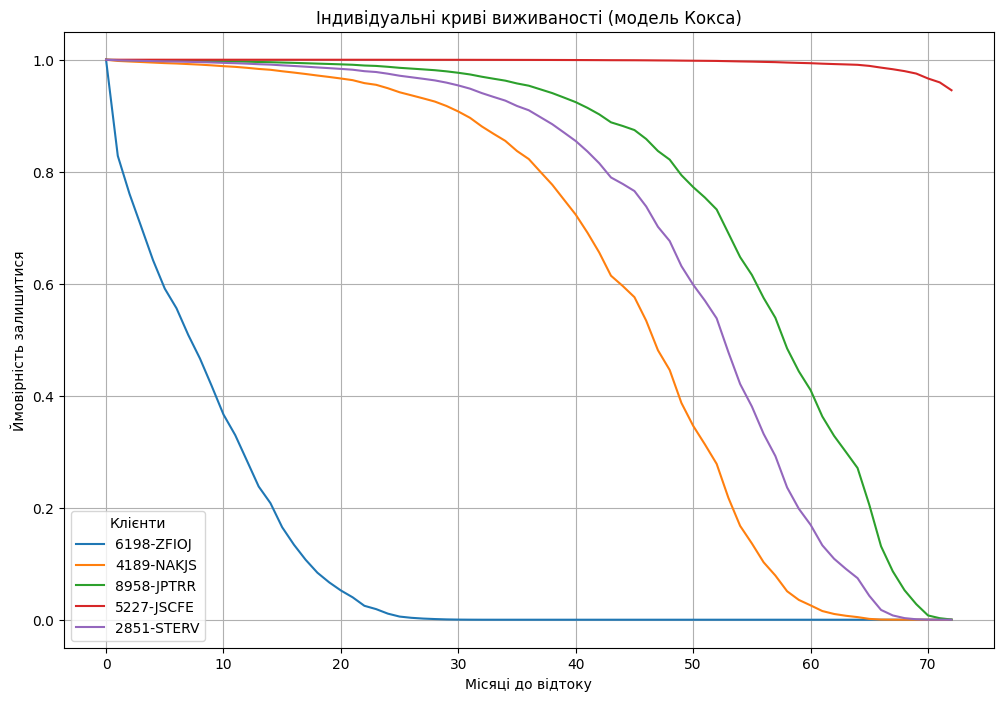
Формула оцінює втрати компанії за одного клієнта, враховуючи:

\* Очікуваний дохід, який клієнт принесе до моменту відтоку.

\* Ризик відтоку клієнта.

Ідея полягає в тому, що клієнт, який має високу ймовірність відтоку, "коштує" компанії більше втрачених доходів, ніж клієнт із низькою ймовірністю відтоку.

Тепер побудуємо графіки:





Результати показують, що більшість клієнтів мають високу ймовірність відтоку (100%), що супроводжується суттєвими прогнозованими збитками. Час до відтоку сильно варіюється: наприклад, для одного клієнта (6198-ZFIOJ) це лише 7 місяців, тоді як для другого (4189-NAKJS) — 46 місяця. Збитки залежать від тривалості часу до відтоку та щомісячних витрат, як видно з різниці між клієнтами 4189-NAKJS ($2298.05) і 2851-STERV ($3691.15).

Четверний Клієнт (5227-JSCFE), з ймовірністю відтоку лише 5.45%, демонструє стабільність, що можна використати для покращення роботи з іншими клієнтами. Криві виживаності підтверджують стрімке зниження ймовірності залишитися для ризикованих клієнтів, що потребує термінових дій.

**Антиризикова стратегія**

Антикризова Стратегія передбачає диференційований підхід до утримання клієнтів залежно від групи ризику. Для клієнтів з високою ймовірністю відтоку (більше 80%) варто розробити комплекс негайних заходів, типу: персональні знижки, індивідуальні пропозиції та розширення пакету послуг. Клієнтам групи середнього ризику пропонувати маркетингові акції, бонусні програми та безкоштовне підвищення рівня обслуговування.

Ключовою особливістю стратегії є врахування типу контракту. Для місячних контрактів потрібно запровадити механізми переведення на довгострокові умови, а для дворічних - надавати максимальні привілеї та ексклюзивні пропозиції. Особливу увагу треба приділити клієнтам, які розраховуються електронним чеком, з розробкою спеціальних утримуючих програм.

Модель базується на точному розрахунку очікуваних втрат для кожного клієнта з диференційованим інвестуванням ресурсів у найцінніші сегменти. Технологічний компонент включає впровадження передбачувальної аналітики та моніторингу ризиків відтоку.

Операційні зміни охоплюють перегляд тарифної політики, оптимізацію комунікаційних каналів та підвищення якості технічної підтримки.

## Висновоки

Проведене дослідження представляє комплексний аналіз проблеми відтоку абонентів, що демонструє критичну важливість попередження втрати клієнтів для сучасного бізнесу. Використані методи машинного навчання, зокрема логістична регресія та модель пропорційних ризиків Кокса, дозволили глибоко дослідити фактори, що впливають на рішення клієнтів залишити компанію.

**Ключові результати дослідження:**

1. Встановлено, що найбільший вплив на відтік мають такі фактори: тривалість перебування (tenure), тип контракту, інтернет-послуги та спосіб оплати.
2. Логістична регресія показала найкращі результати Accuracy: 0.8027 F1 Score: 0.5983 , що підтверджує її ефективність у прогнозуванні відтоку.
3. Виявлено критичну різницю в лояльності клієнтів залежно від типу контракту: дворічні контракти значно знижують ризик відтоку порівняно з місячними.
4. Розроблена антикризова стратегія передбачає диференційований підхід до утримання клієнтів з урахуванням індивідуальних ризиків та економічної цінності.

Практичне значення дослідження полягає в можливості завчасно ідентифікувати групи ризику та впроваджувати цільові заходи утримання, що потенційно дозволить знизити рівень відтоку та покращити фінансові показники компанії.

Перспективи подальших досліджень включають вдосконалення прогностичних моделей, поглиблений аналіз нематеріальних факторів впливу та розробку більш складних стратегій утримання клієнтів.